What is a Data Commons and Why Should You Care?

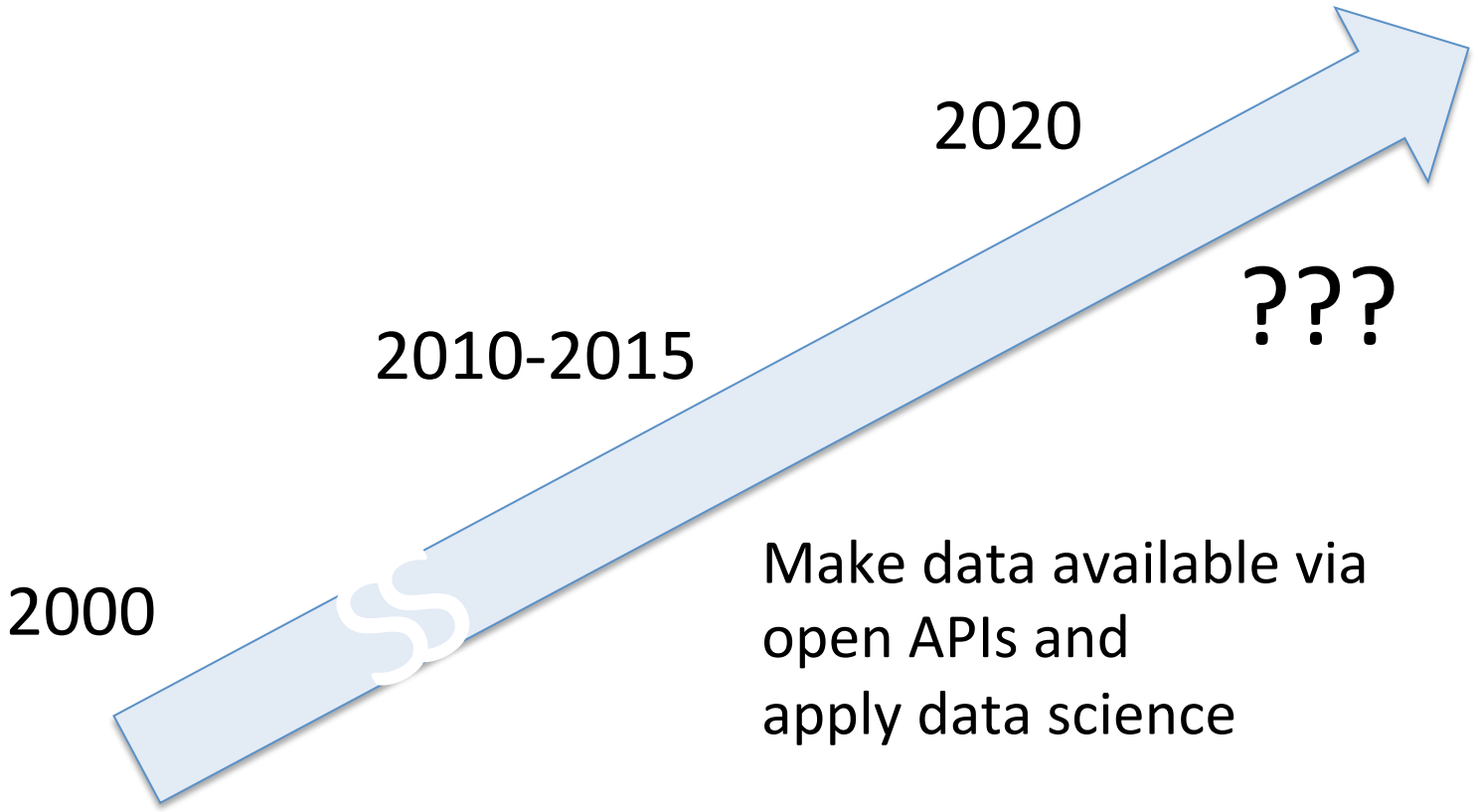
Robert Grossman

University of Chicago

Open Cloud Consortium

April 22, 2015

NASA IS&T Colloquium


2020
???
2010-2015
2000
Make data available via
open APIs and
apply data science
Collect data and
distribute files via DAAC
and apply data mining

1. Data Commons

We have a problem …

The commoditization of sensors is creating an explosive growth of data

There is not enough funding for every researcher to house all the data they need

It can take weeks to download large geo-spatial datasets

Analyzing the data is more expensive than producing it

Data Commons

Data commons co-locate data, storage and computing infrastructure, and commonly used tools for analyzing and sharing data to create a resource for the research community.

Source: Interior of one of Google's Data Center, www.google.com/about/datacenters/

The Tragedy of the Commons

Garrett Hardin

Individuals when they act independently following their self interests can deplete a deplete a common resource, contrary to a whole group's long-term best interests.

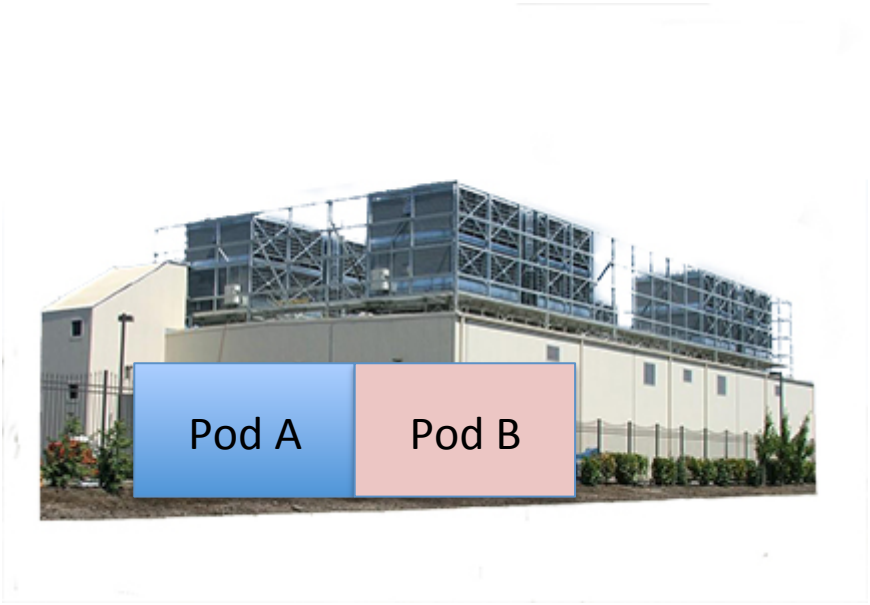
Source: Garrett Hardin, The Tragedy of the Commons, Science, Volume 162, Number 3859, pages 1243-1248, 13 December 1968.

- U.S based not-for-profit corporation.
- Manages cloud computing infrastructure to support scientific research: Open Science Data Cloud, Project Matsu, & OCC NOAA Data Commons.
- Manages cloud computing infrastructure to support medical and health care research: Biomedical Commons Cloud.
- Manages cloud computing testbeds: Open Cloud Testbed.

www.opencloudconsortium.org

What Scale?

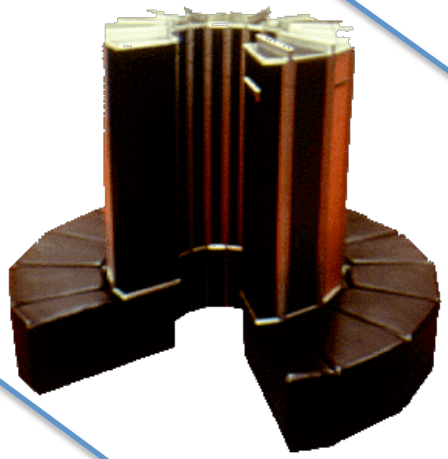
- New data centers are sometimes divided into "pods," which can be built out as needed.
- A reasonable scale for what is needed for a commons is one of these pods ("cyberpod")
- Let's use the term "datapod" for the analytic infrastructure that scales to a cyberpod.
- Think of as the scale out of a database.



2004
10x-100x

data science

1976
10x-100x



simulation
science

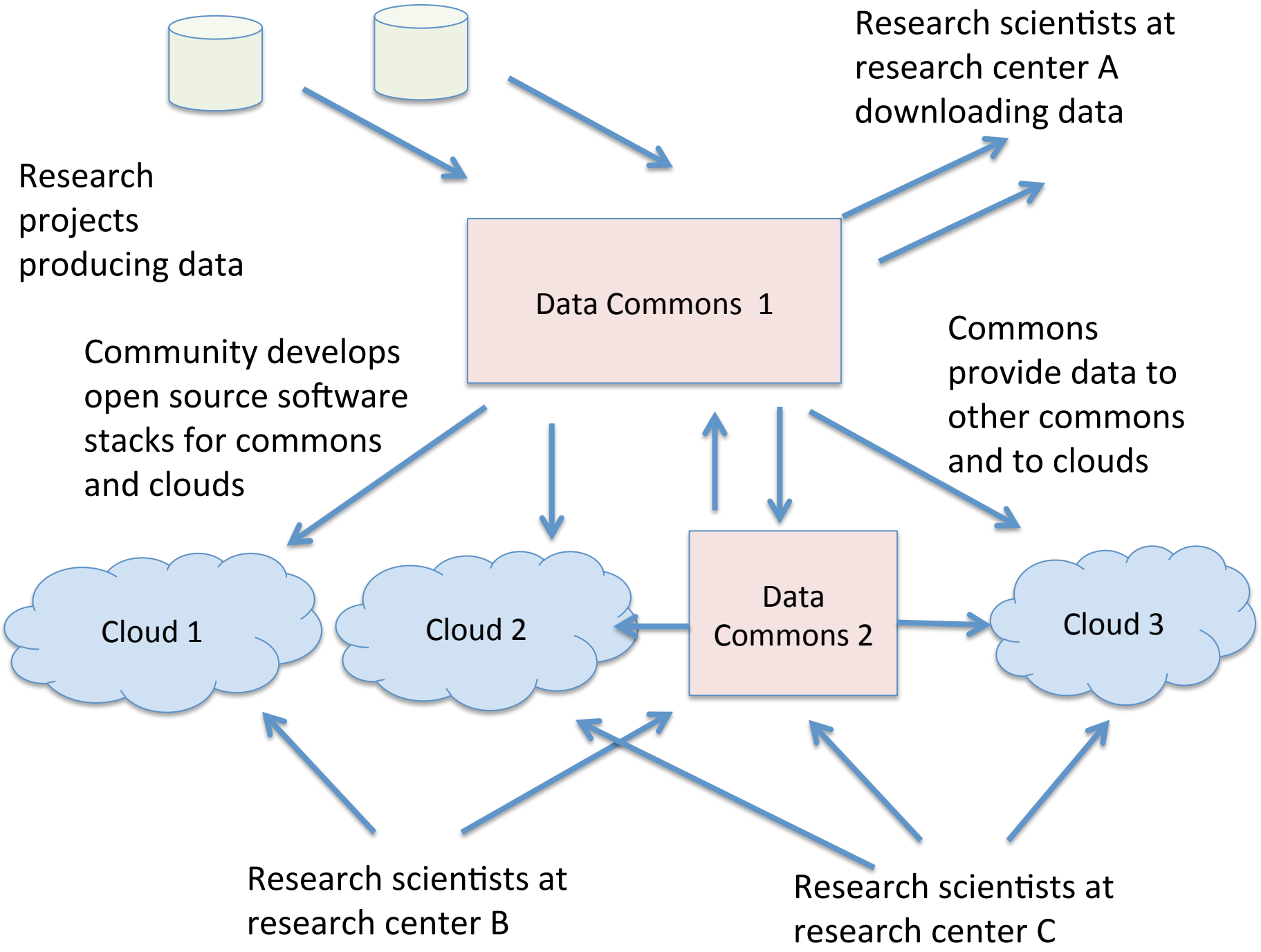
1670
250x

1609
30x

experimental
science

Core Data Commons Services

- Digital IDs
- Metadata services
- High performance transport
- Data export
- Pay for compute with images/containers containing commonly used tools, applications and services, specialized for each research community


Research scientists at research center A downloading data
Research projects producing data
Data Commons 1
Community develops open source software stacks for commons and clouds
Commons provide data to other commons and to clouds
Cloud 1
Cloud 2
Data Commons 2
Cloud 3
Research scientists at research center B
Research scientists at research center C

Characteristic	Colectomy		Gastrectomy	
	Patients (%)	Mortality rate (%)	Patients (%)	Mortality rate (%)
Age > 65	28,243 (58.2)	6.7*	3,482 (54.1)	2.5*
Female gender	26,257 (54.1)	4.8‡	2,987 (46.4)	8.4
African American	4,553 (9.4)	5.0	915 (14.2)	8.5
Medicaid	2,843 (5.9)	4.7	659 (10.2)	6.2‡
IHD	7,255 (14.9)	8.6*	942 (14.6)	13.6*
Airway obstruction	1,782 (3.7)	3.7	271 (4.2)	7.0
CHD	4,335 (8.9)	16.8*	613 (9.5)	24.5*
Metastasis	5,953 (12.3)	6.6*	1,099 (17.1)	9.0
PVD	265 (0.6)	18.9*	46 (0.7)	21.7†
COPD	4,004 (8.2)	9.9*	556 (8.6)	16.4*
Diabetes	6,703 (13.8)	6.2*	975 (15.2)	9.0
Dysrhythmia	6,464 (13.3)	14.7*	987 (15.3)	22.9*
All patients	48,582 (100)	4.6	6,434 (100)	8.4

CHD indicates congestive heart disease; COPD, chronic obstructive pulmonary disease; IHD, ischemic heart disease; PVD, peripheral vascular disease.

cyber pods

memory databases datapods

GB TB PB

W KW MW

Complex statistical models over small data that are highly manual and update infrequently.

Simpler statistical models over large data that are highly automated and update frequently.

Is More Different? Do New Phenomena Emerge at Scale in Biomedical Data?

4 August 1972, Volume 177, Number 4047 **SCIENCE**

More Is Different

Broken symmetry and the nature of the hierarchical structure of science.

P. W. Anderson

The reductionist hypothesis may still be a topic for controversy among philosophers, but among the great majority of active scientists I think it is accepted without question. The workings of our minds and bodies, and of all the animate or inanimate matter of which we

planation of phenomena in terms of known fundamental laws. As always, distinctions of this kind are not unambiguous, but they are clear in most cases. Solid state physics, plasma physics, and perhaps also biology are extensive. High energy physics and a good part of nuclear physics are intensive. There is always much less intensive research going on than extensive.

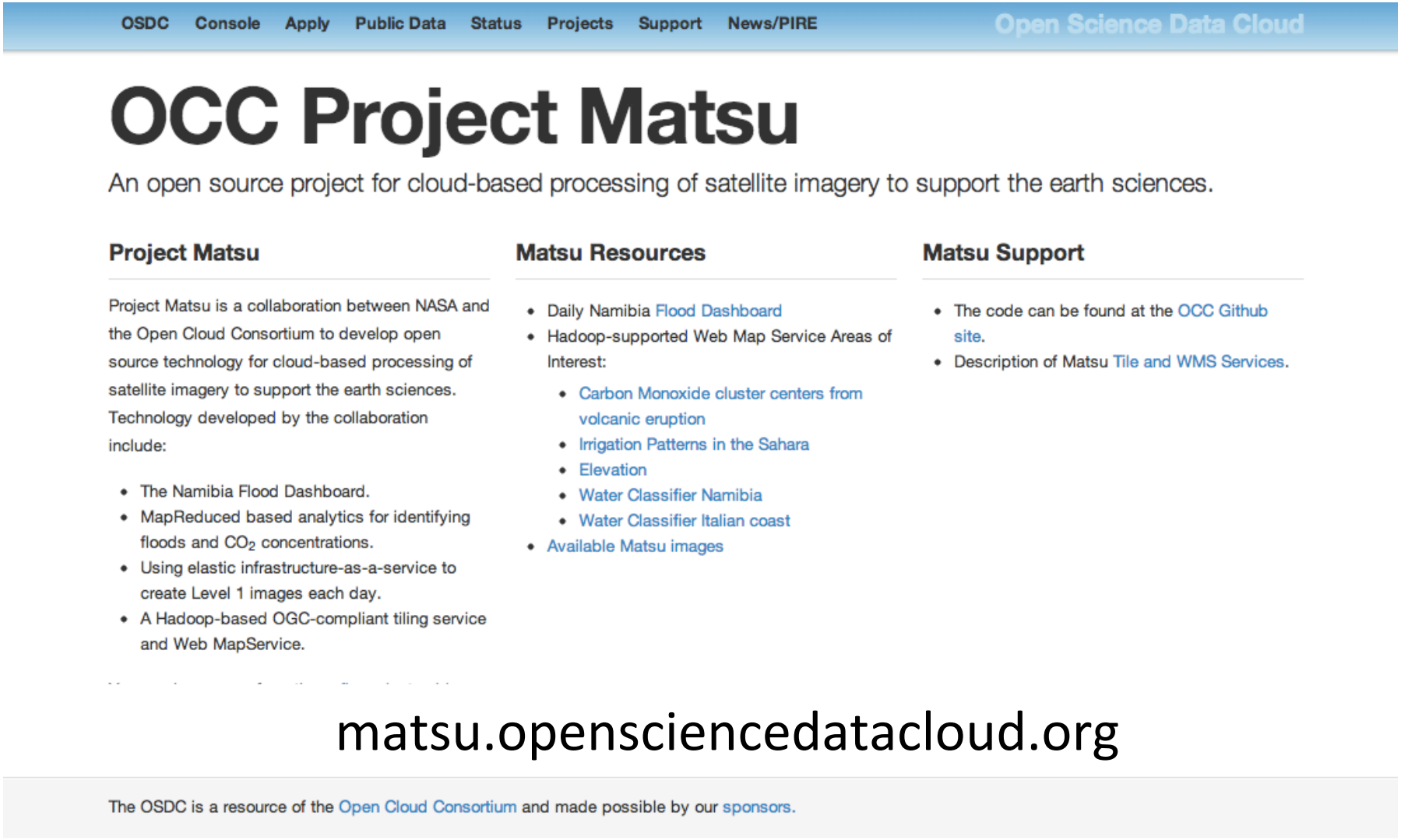
less relevance they seem to have to the very real problems of the rest of science, much less to those of society.

The constructionist hypothesis breaks down when confronted with the twin difficulties of scale and complexity. The behavior of large and complex aggregates of elementary particles, it turns out, is not to be understood in terms of a simple extrapolation of the properties of a few particles. Instead, at each level of complexity entirely new properties appear, and the understanding of the new behaviors requires research which I think is as fundamental in its nature as any other. That is, it seems to me that one may array the sciences roughly linearly in a hierarchy, according to the idea: The elementary entities of science X obey the laws of science Y.

X	Y
solid state or many-body physics	elementary particle physics

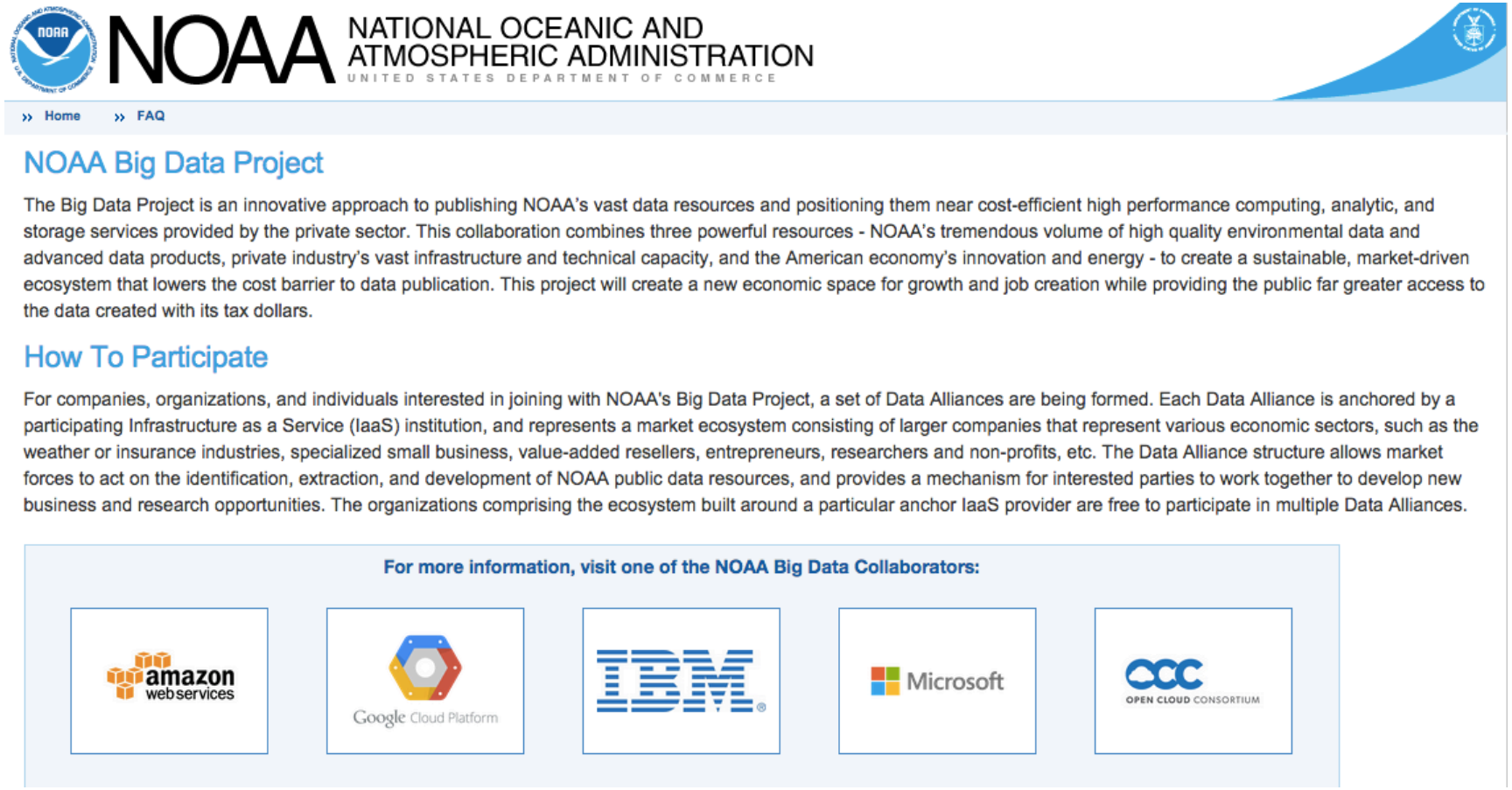
Source: P. W. Anderson, More is Different, Science, Volume 177, Number 4047, 4 August 1972, pages 393-396.

2. OCC Data Commons



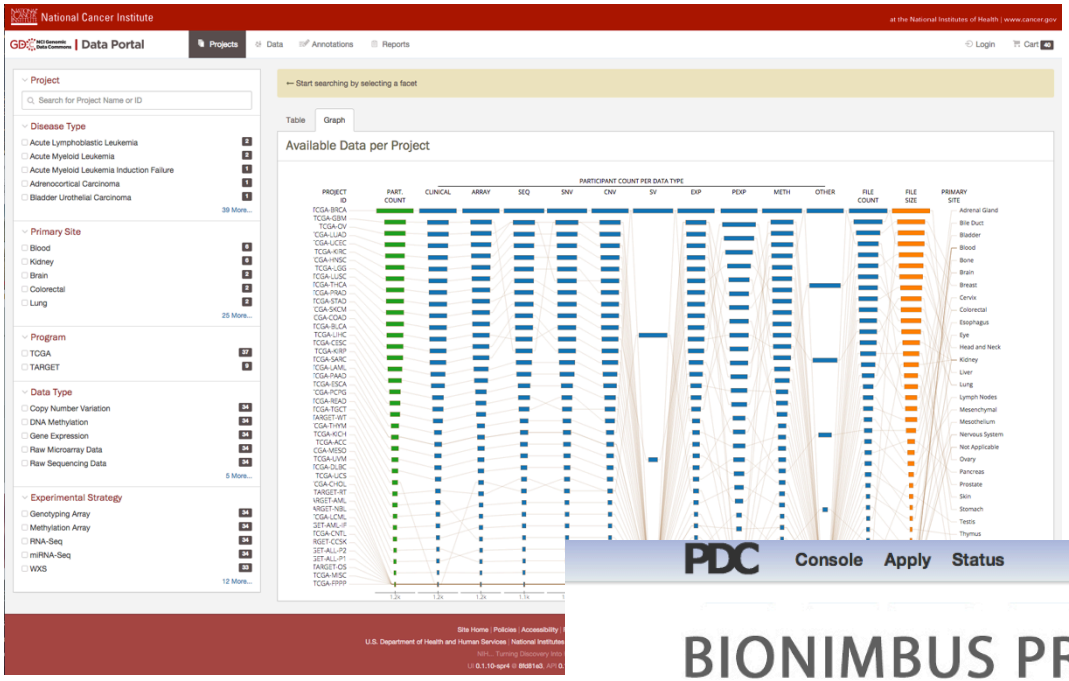
matsu.opensciencedatacloud.org

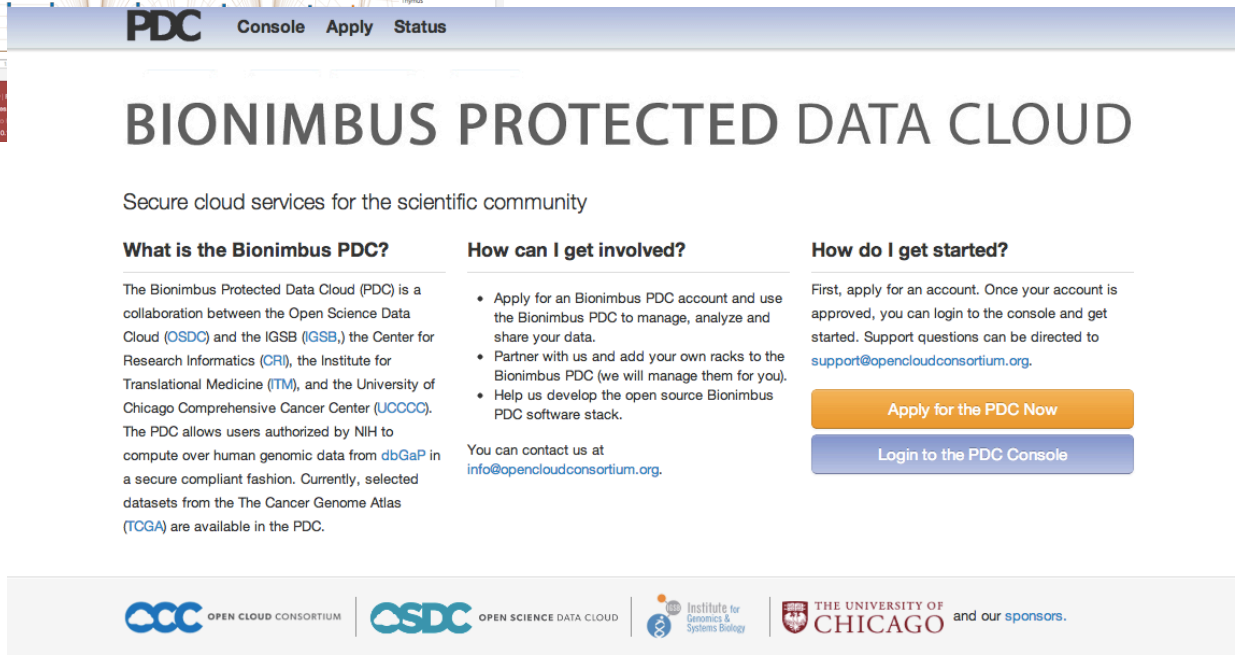
OCC-NASA Collaboration 2009 - present



- Public-private data collaborative announced April 21, 2015 by Secretary of Commerce Pritzker.
- AWS, Google, Microsoft and Open Cloud Consortium will form four collaborations.

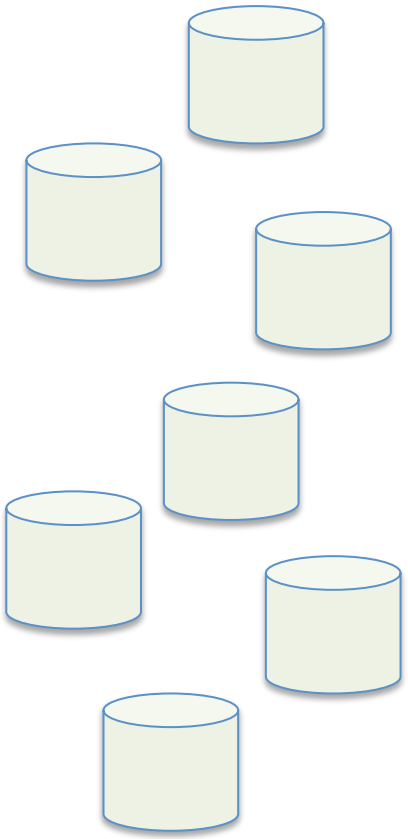
NCI Genomic Data Commons



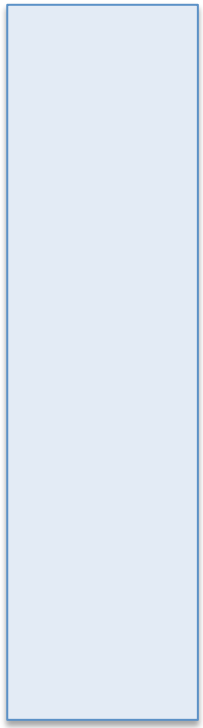


OSDC Commons Architecture

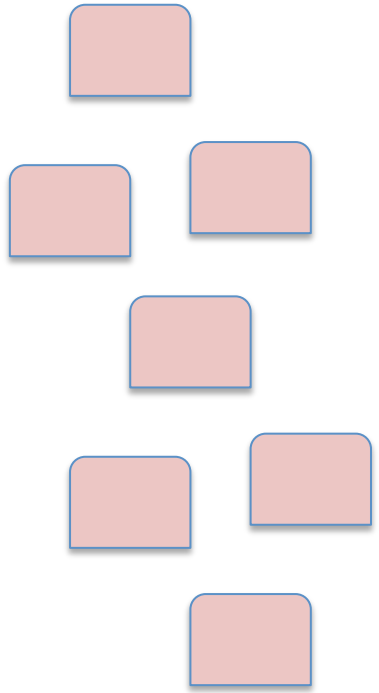
Digital ID Service
& Metadata Service



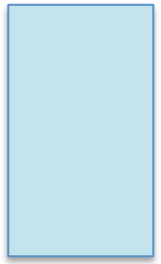
Object storage
(permanent)



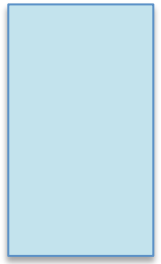
Scalable light
weight workflow



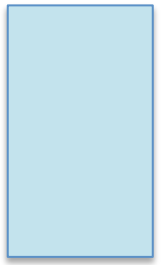
Community data products
(data harmonization)



Co-located
“pay for
compute”



Open APIs
for data
access and
data access
portal



Data
submission
portal

Devops supporting virtual machines and containers

3. Scanning Queries over Commons and the Matsu Wheel

What is the Project Matsu?

Matsu is an open source project for processing satellite imagery to support earth sciences researchers using a data commons.

Matsu is a joint project between the Open Cloud Consortium and NASA's EO-1 Mission (Dan Mandl, Lead)

All available L1G images (2010-now)

NASA's Matsu Mashup

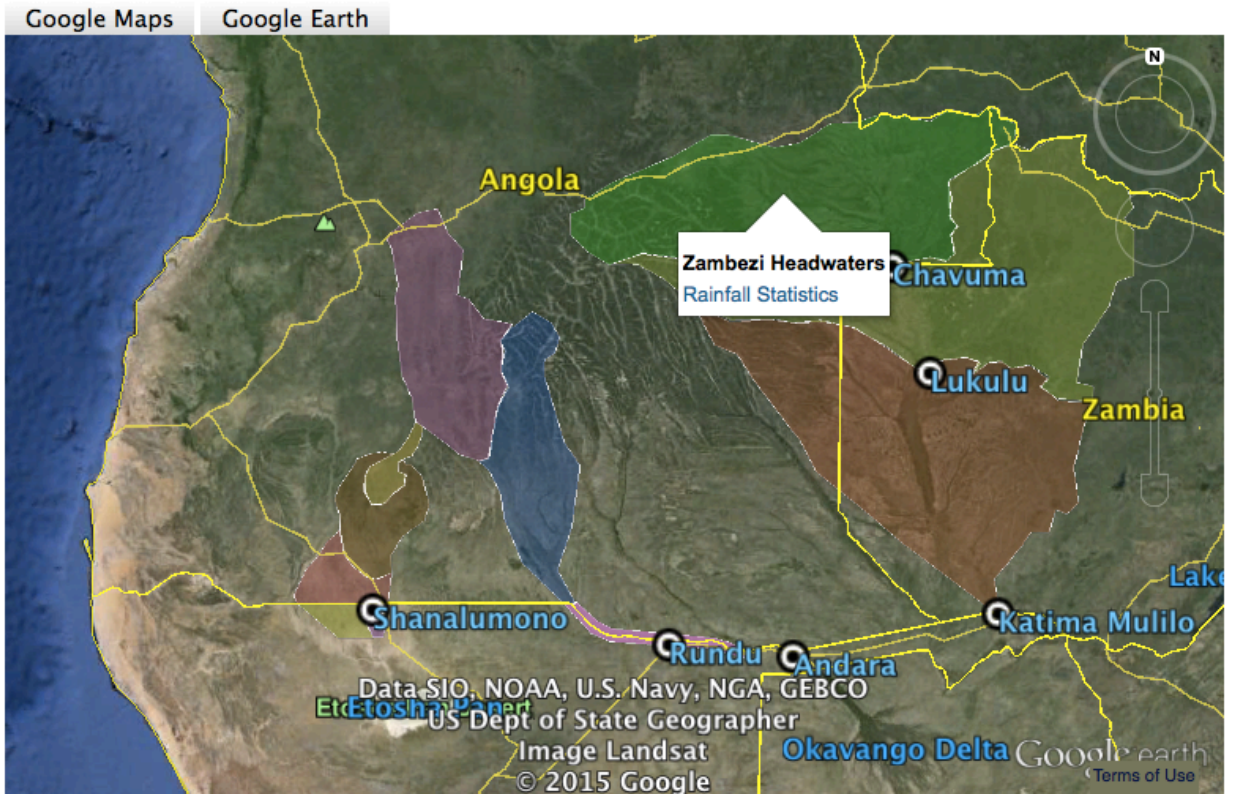
April
22

View today's Daily Report

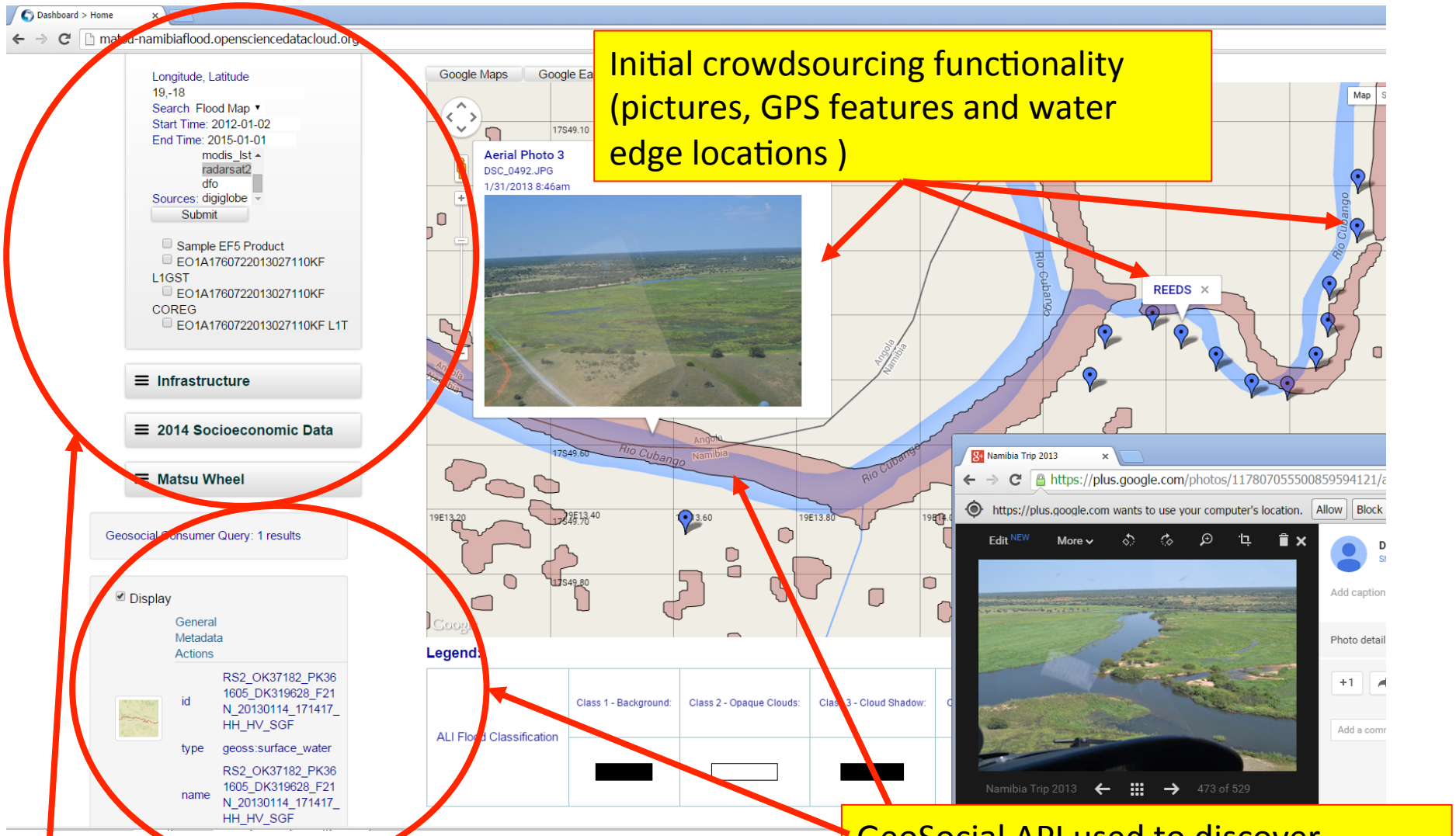
Namibia Flood Dashboard

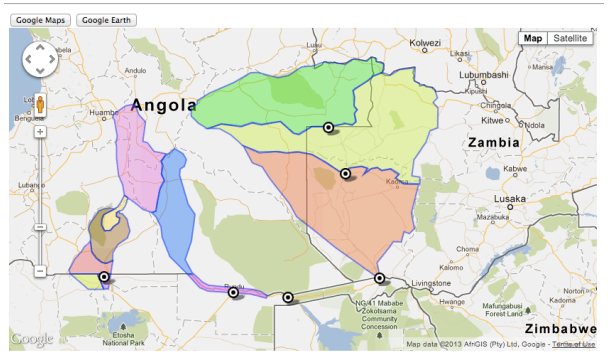
New Bulletin | View Current Bulletin | View Bulletin Records

- SensorWeb Layers
- Water Lines and Areas
- Satellite Overlays
- Ground Pics
- Kavango Radarsat Data
- Cuvelai Radarsat Data
- TRMM Rainfall Accumulation and Flood Forecast
- Global Scene Counts



Flood/Drought Dashboard Examples



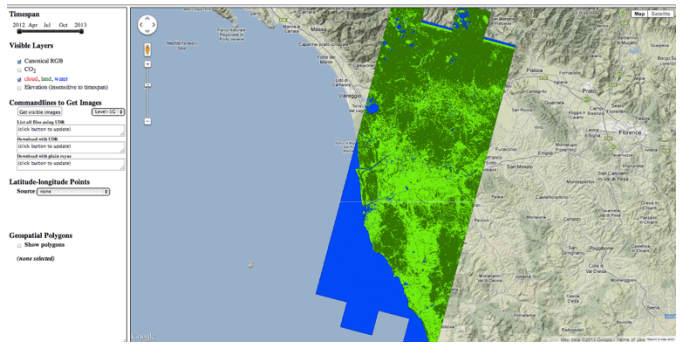


Legend:

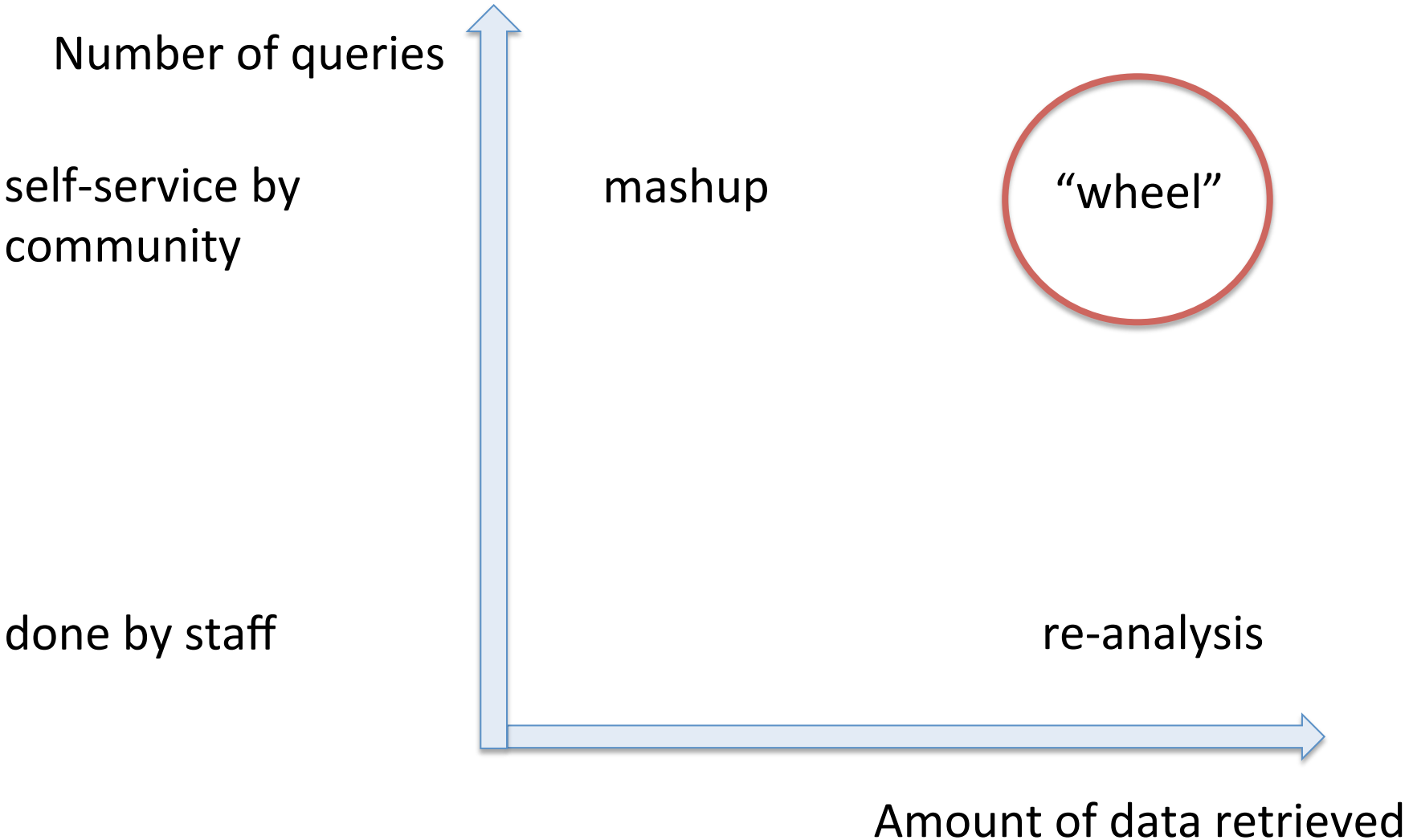
ALI Flood Classification	Class 1 - Background:	Class 2 - Opaque Clouds:	Class 3 - Cloud Shadow:	Class 4 - Haze and Thin Clouds:	Class 5 - Clear Water:	Class 6 - Turbid Water:	Class 7 - Dry Land:

1. Open Science Data Cloud (OSDC) stores Level 0 data from EO-1 and uses an OpenStack-based cloud to create Level 1 data.

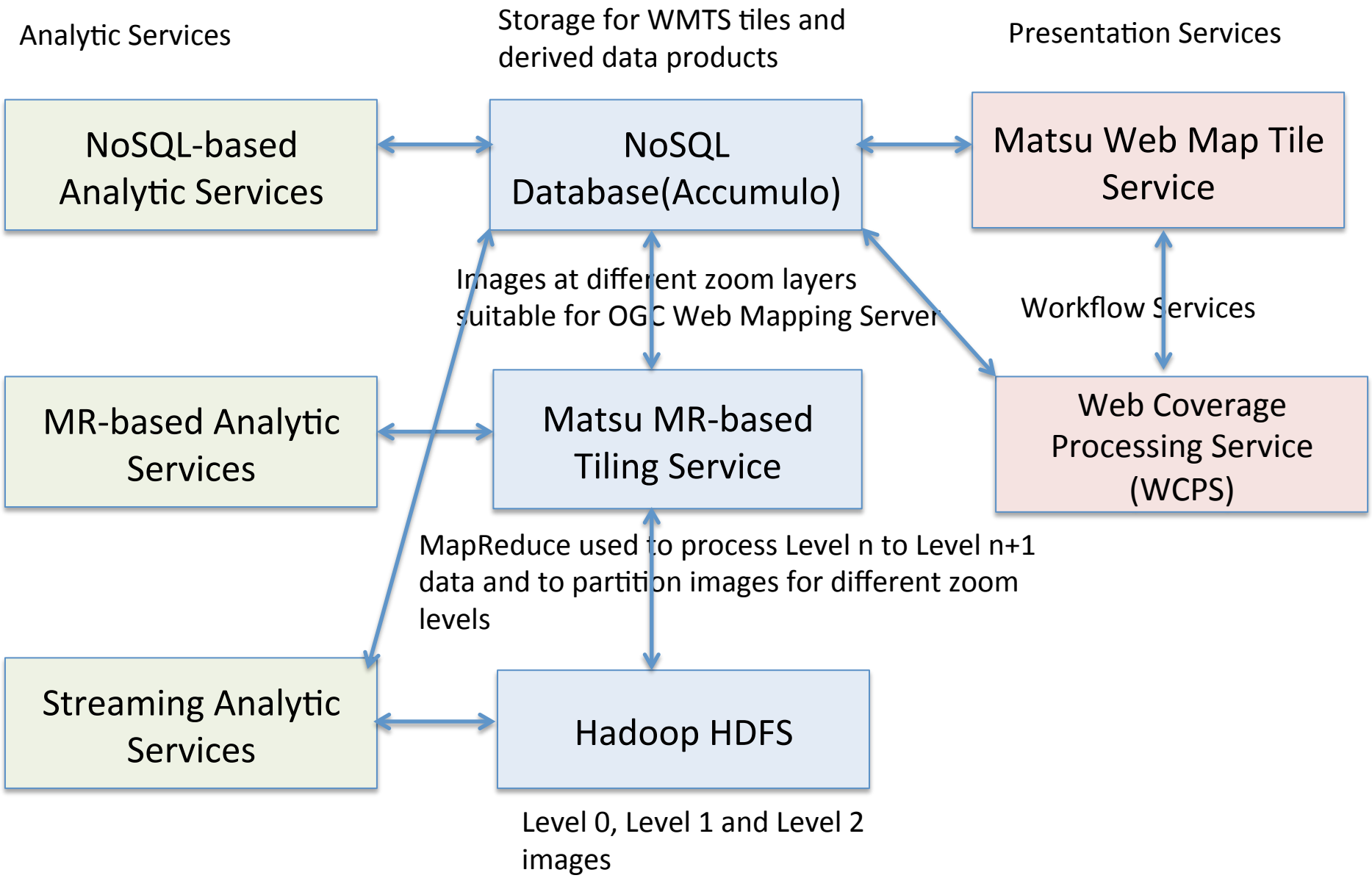
2. OSDC also provides OpenStack resources for the Nambia Flood Dashboard developed by Dan Mandl's team.



3. Project Matsu uses a Hadoop/Accumulo system to run analytics nightly and to create tiles with OGC-compliant WMTS.

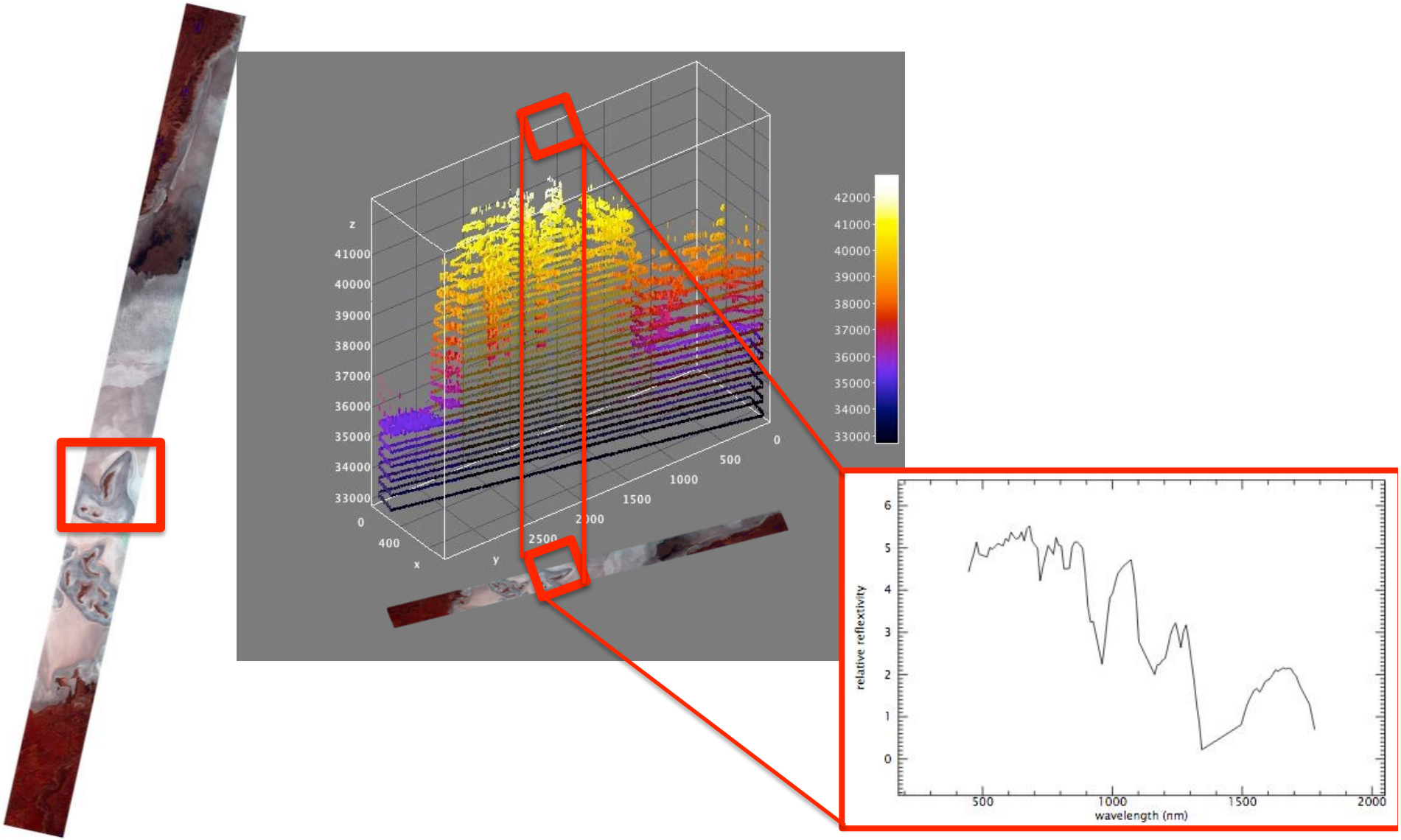

Number of queries
self-service by community
done by staff
mashup
"wheel"
re-analysis
Amount of data retrieved
row-oriented
column-oriented

Matsu Hadoop Architecture

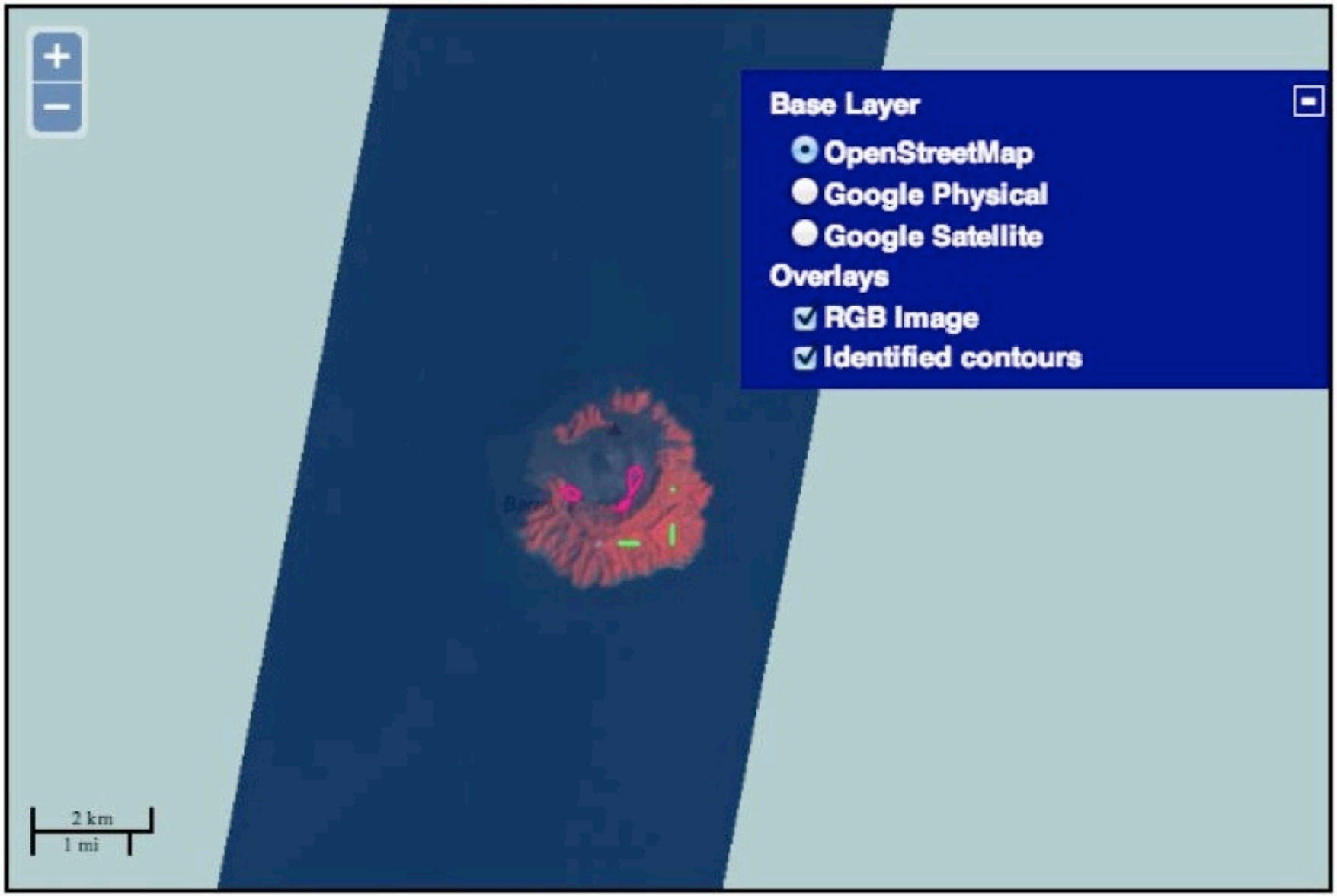


The Matsu Wheel

for analyzing large volumes of hyperspectral image data



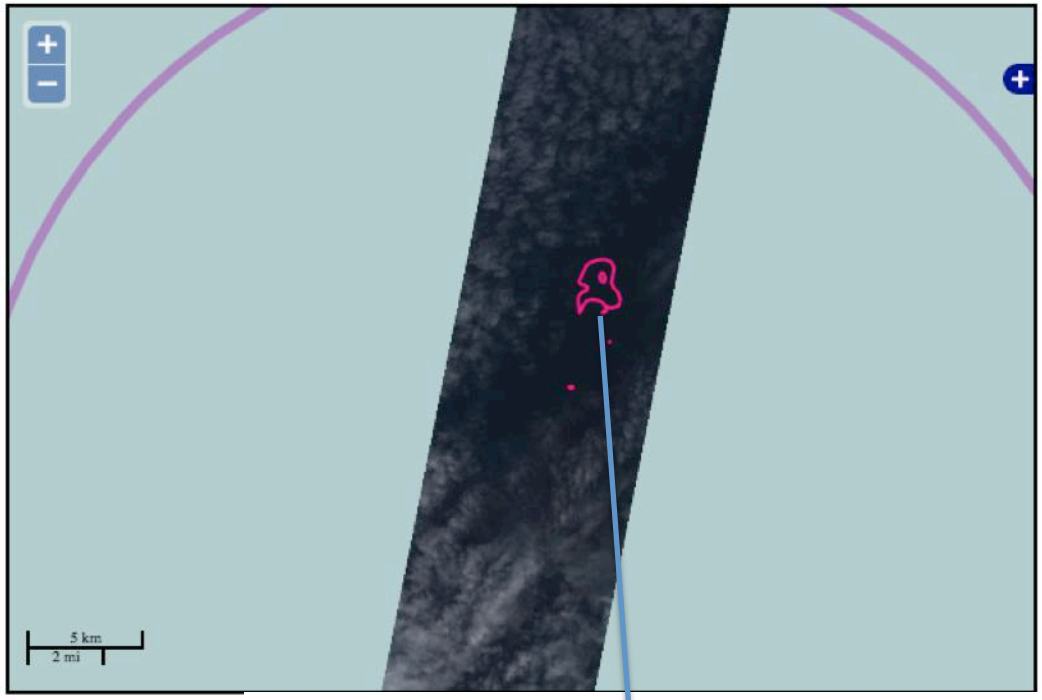
Spectral anomaly detected:
Barren Island active volcano, Feb, 2014



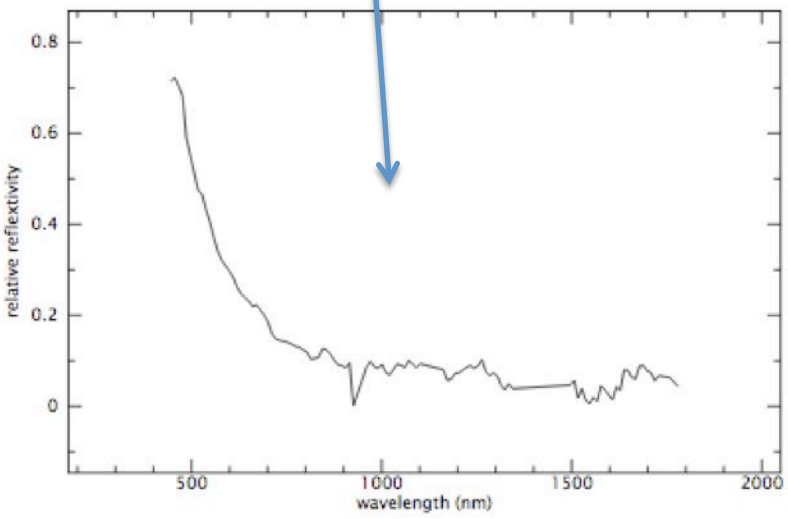
Spectral anomaly detected: Nishinoshima active volcano, Dec, 2014

Matsu Analytic Image Report

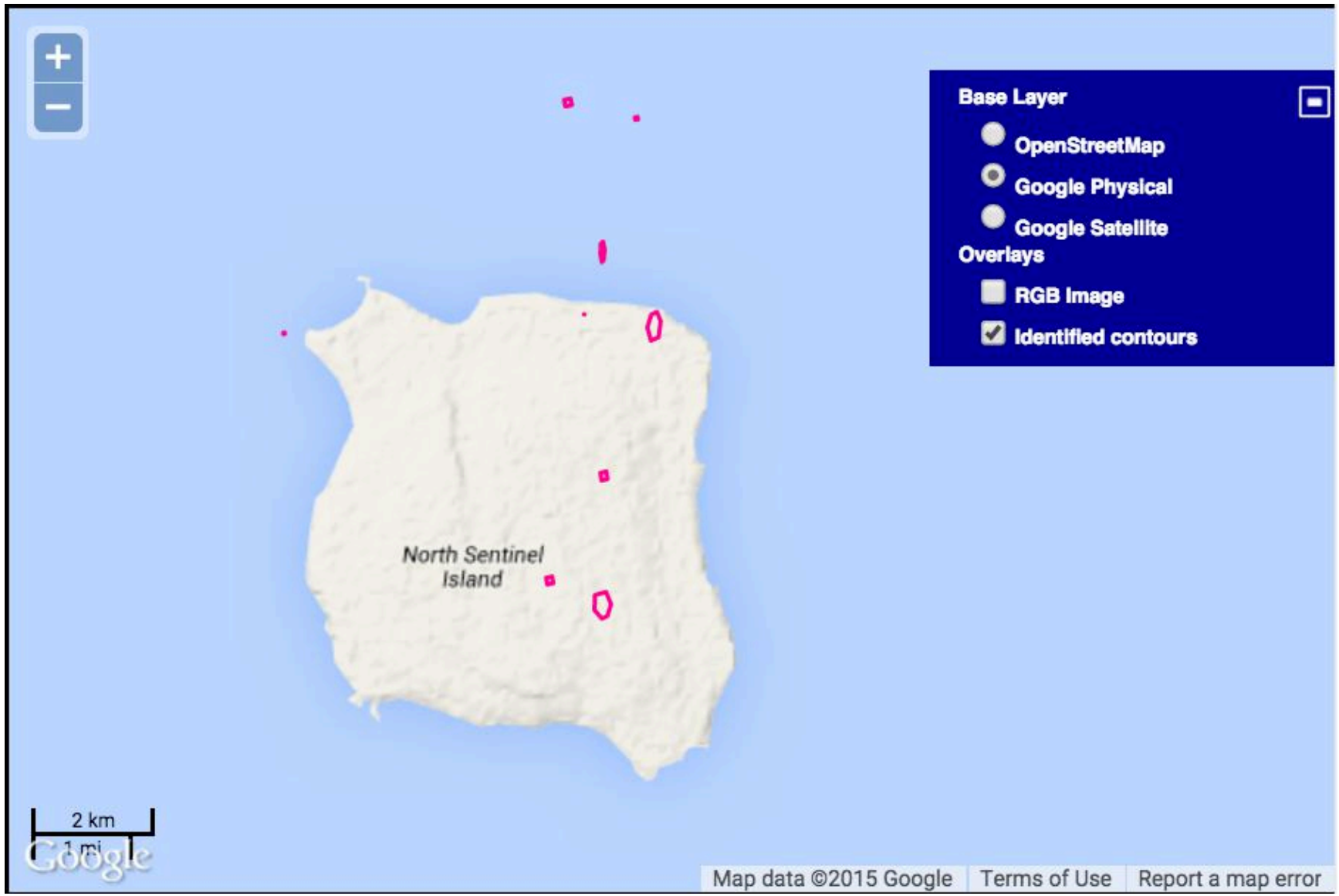
Collection Date	2014-12-02 (day 336)
Analysis Date	Wed Dec 17 12:27:25 2014
Analytic Environment	
Analytic	Contours-2013-12-r4
Noise Correction Enabled	False
Summary Stats	ss-2013-12-r1
Data Ingest	populateHDFS-2013-11-r1
Report Format	reportContoursR4
Hyperspectral Image	
Image	EO1H1050412014336110KF_HYP_L1G
Number of Bands	242



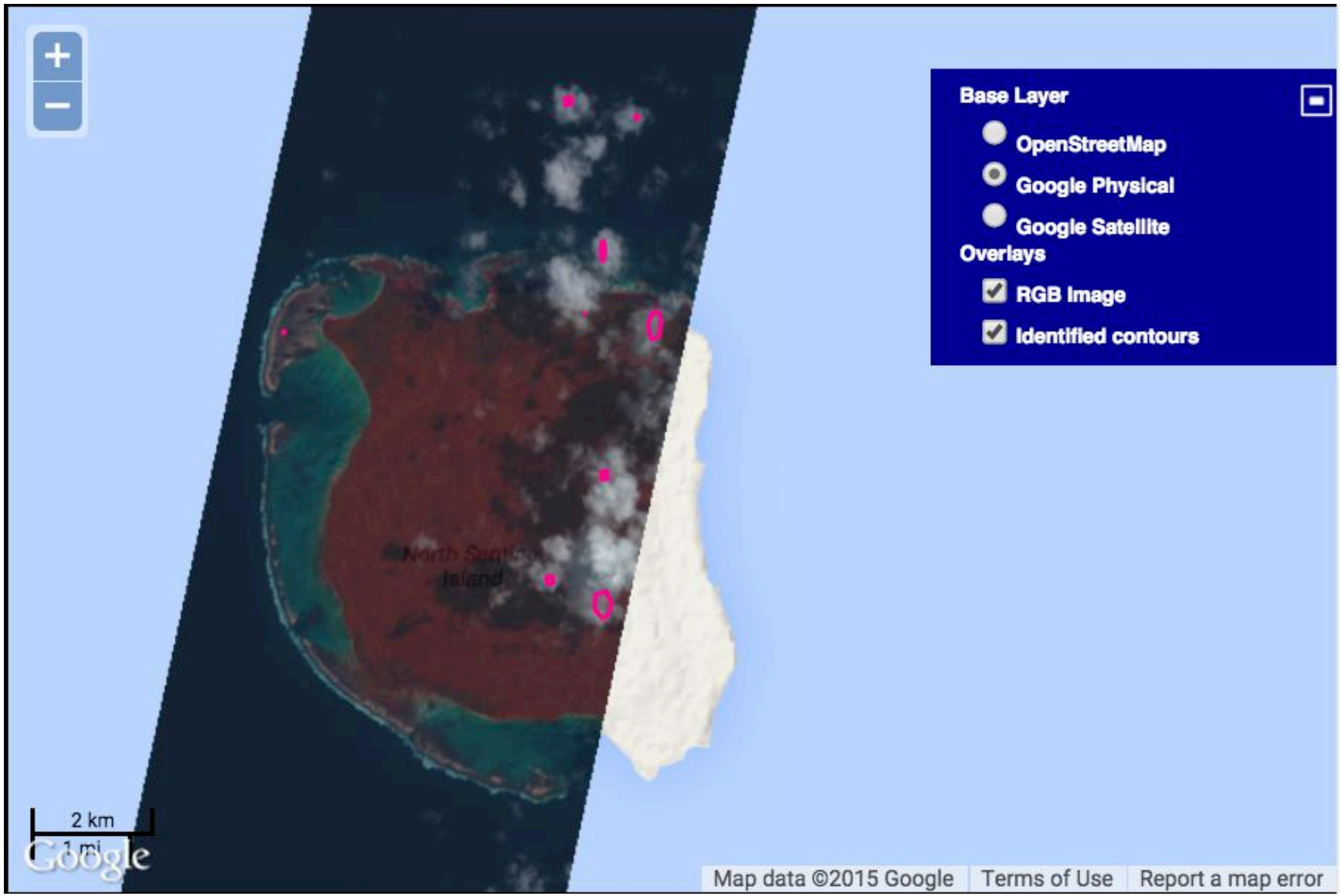
Contour ID	Cluster Score	Contour Score	lat,lng	Area (Pixels)	Area (Meters)	color
C1-05041-0KF	351	0.9719	140.886733625,27.2918559268	7.9589	6259.0137	COLOR
C1-05041-0KF	351	1.0807	140.897972808,27.3285963336	2447.4154	1925311.5337	COLOR
C1-05041-0KF	351	1.1266	140.899385769,27.3310296144	66.3332	52183.5335	COLOR
C1-05041-0KF	351	1.4893	140.900233529,27.3190516554	8.5744	6744.6581	COLOR
C1-05041-0KF	351	0.9264	140.902293378,27.3081518463	0.6165	484.8863	COLOR



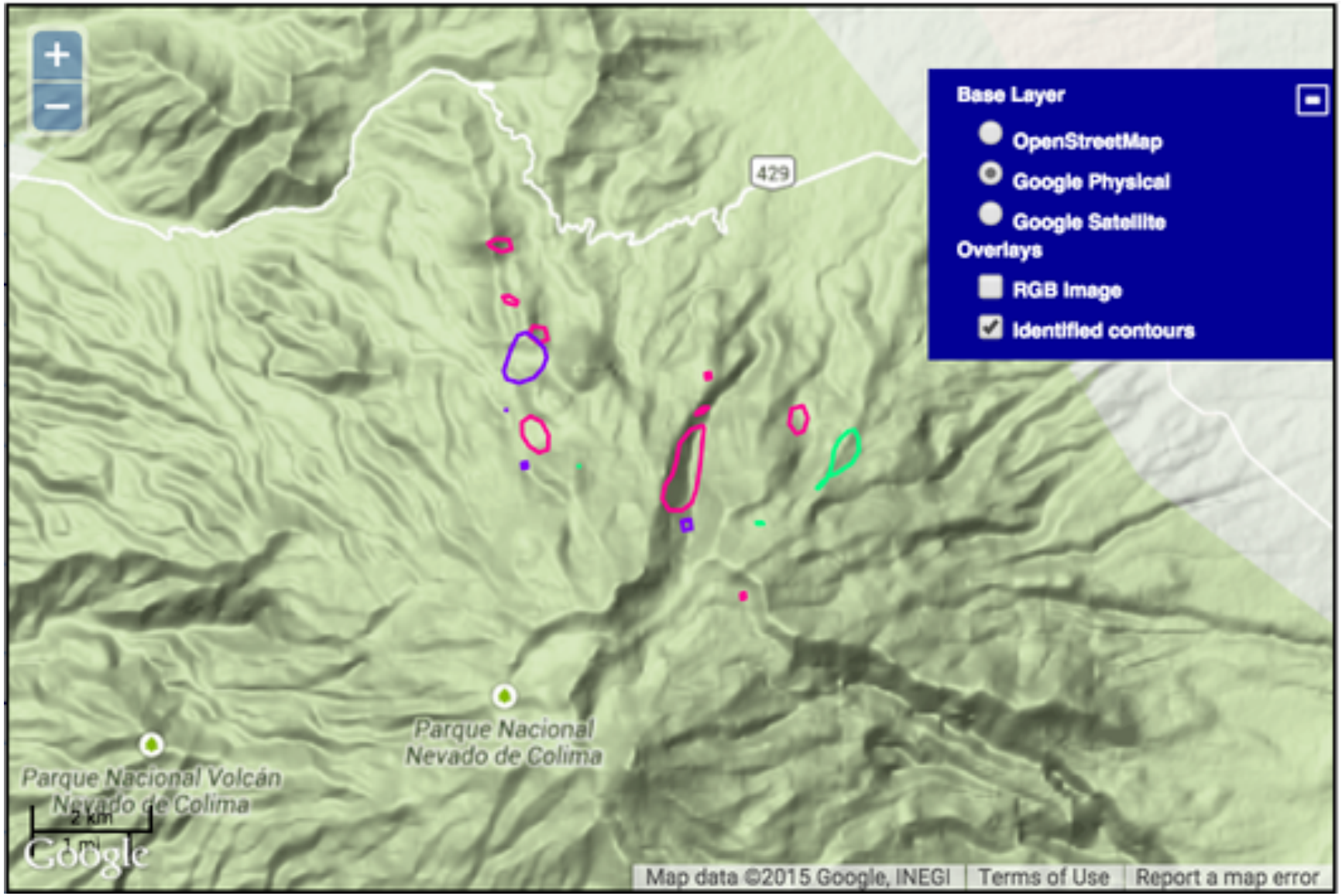
Spectral anomaly detected:
North Sentinel Island fires, May, 2014



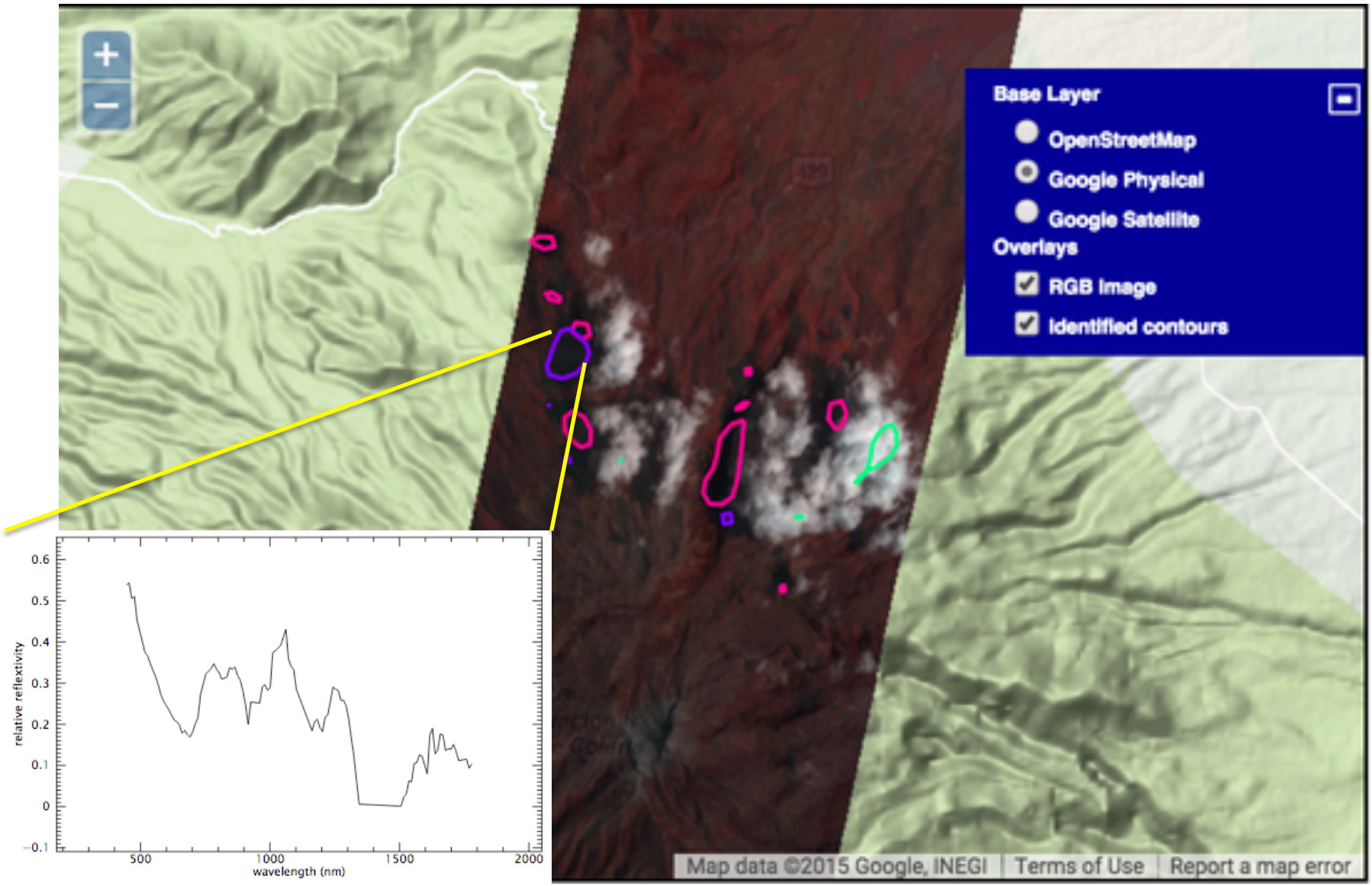
Spectral anomaly detected:
North Sentinel Island fires, May, 2014



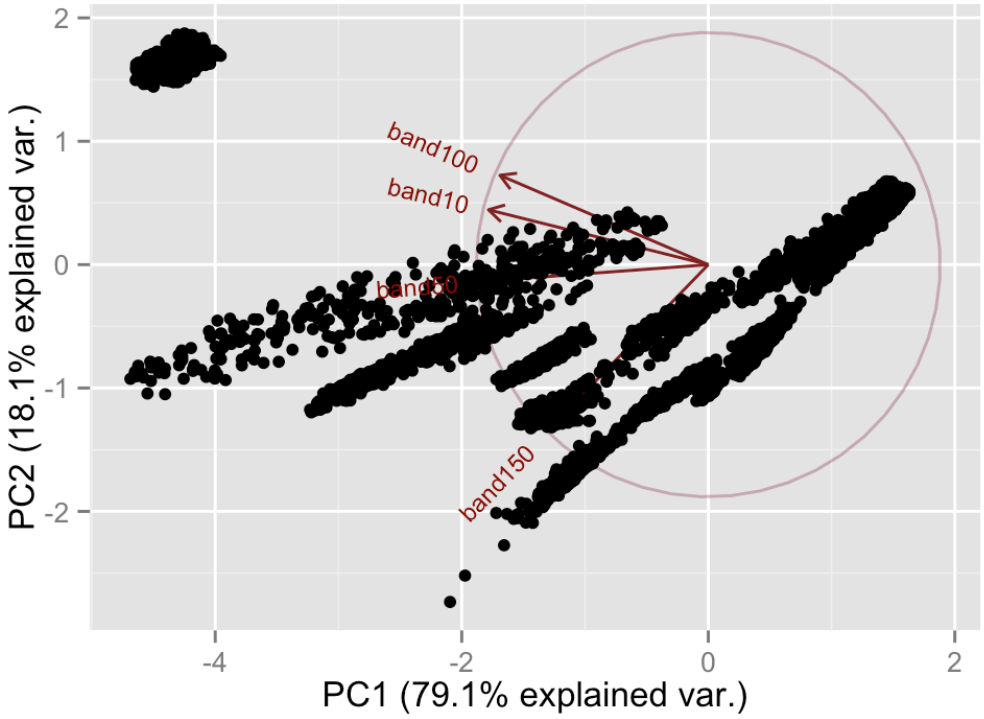
Spectral anomaly detected:
Colima Volcano, April 14, 2015



Spectral anomaly detected: Colima Volcano, April 14, 2015



Matsu Wheel Spectral Anomaly Detector



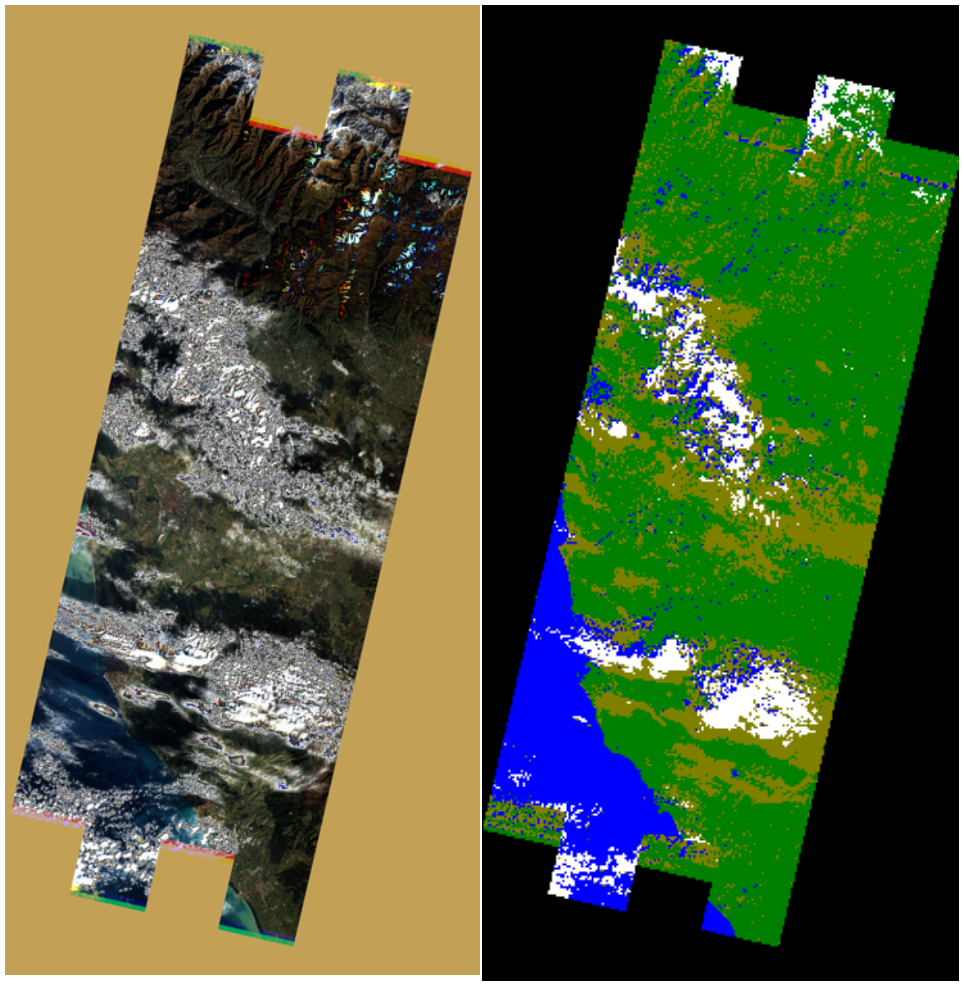
- “Contours and Clusters” – looks for physical contours around spectral clusters
- PCA analysis applied to the set of reflectivity values (spectra) for every pixel, and the top 5 components are extracted for further analysis.

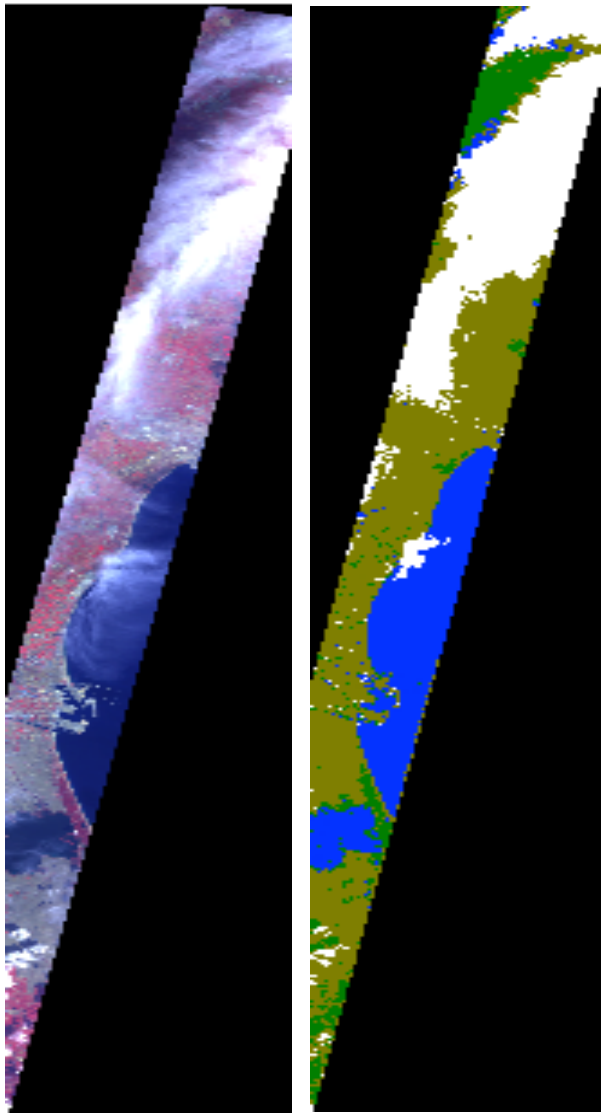
Matsu Wheel Spectral Anomaly Detector

- “Contours and Clusters” – looks for physical contours around spectral clusters
- PCA analysis applied to the set of reflectivity values (spectra) for every pixel, and the top 5 components are extracted for further analysis.
- Pixels are clustered in the transformed 5-D spectral space using a k-means clustering algorithm.
- For each image, k = 50 spectral clusters are formed and ranked from most to least extreme using the Mahalanobis distance of the cluster from the spectral center.
- For each spectral cluster, adjacent pixels are grouped together into contiguous objects.

→ returns geographic regions of spectral anomalies that are scored again as anomalous (0 least , 1000 most) compared to a set of “normal” spectra, constructed for comparison over a baseline of time

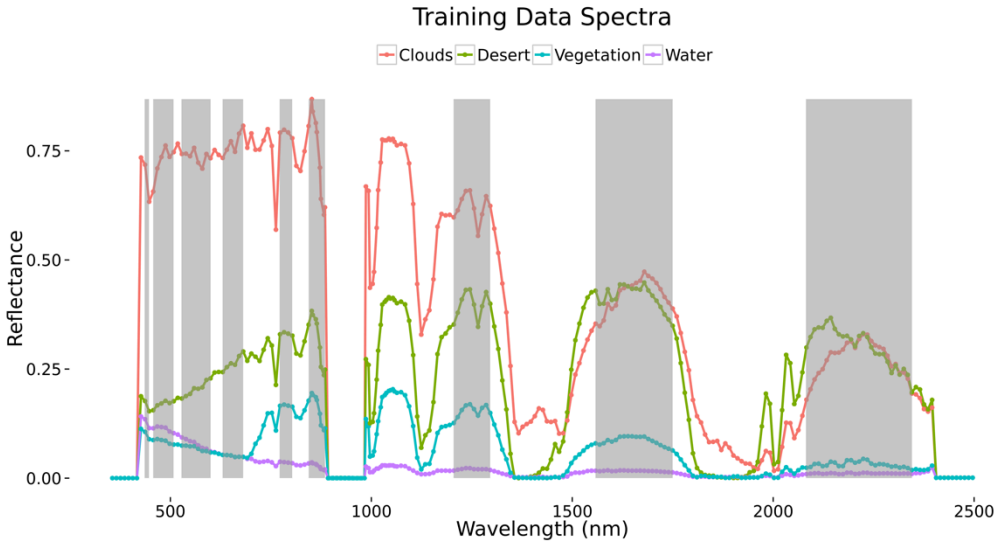
Wheel analytic (beta): SVM-based land cover classifier

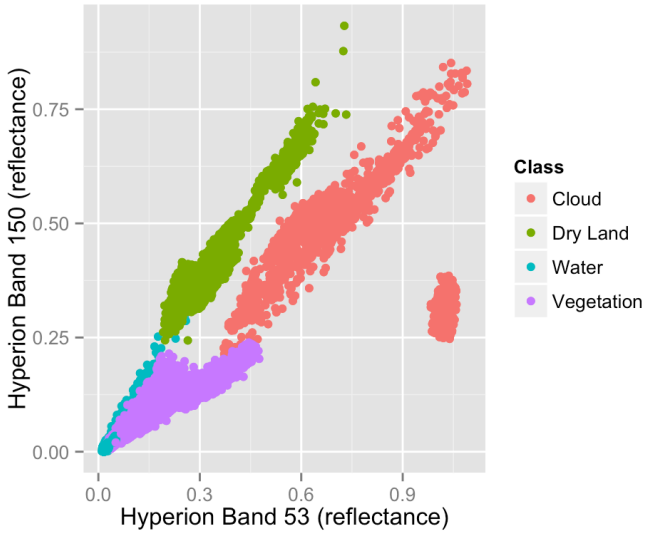




Matsu Wheel Land Cover Classifier

- Support Vector Machine supervised classifier (uses Python's Sci-kit learn)
  - "Training set" constructed from a variety of scenes with range of locations, time of year, sun angle
  - Cloud, Desert/ Dry Land, Water, Vegetation are manually (visually) classified using RGB image





→ returns classified image

4. Data Peering

Hierarchy of the Global Internet

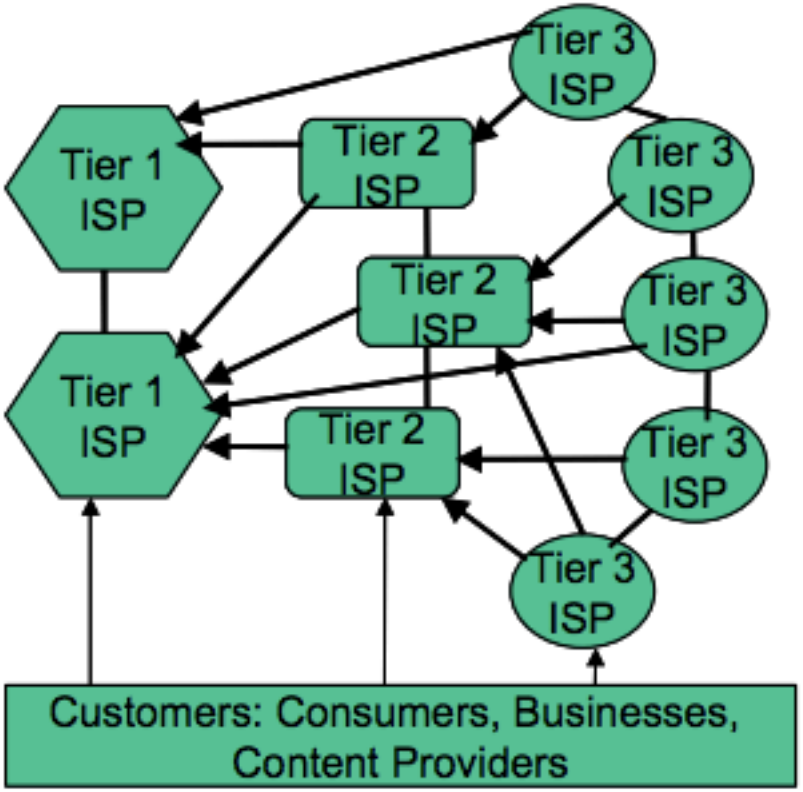
Peering is done between equivalent-sized partners (tier 3 to tier 3).

Transit or fee-based peering is done where there are unequal traffic flows (tier 2 to tier 1).

Peering and transit arrangements may be established directly or at third-party exchange points.

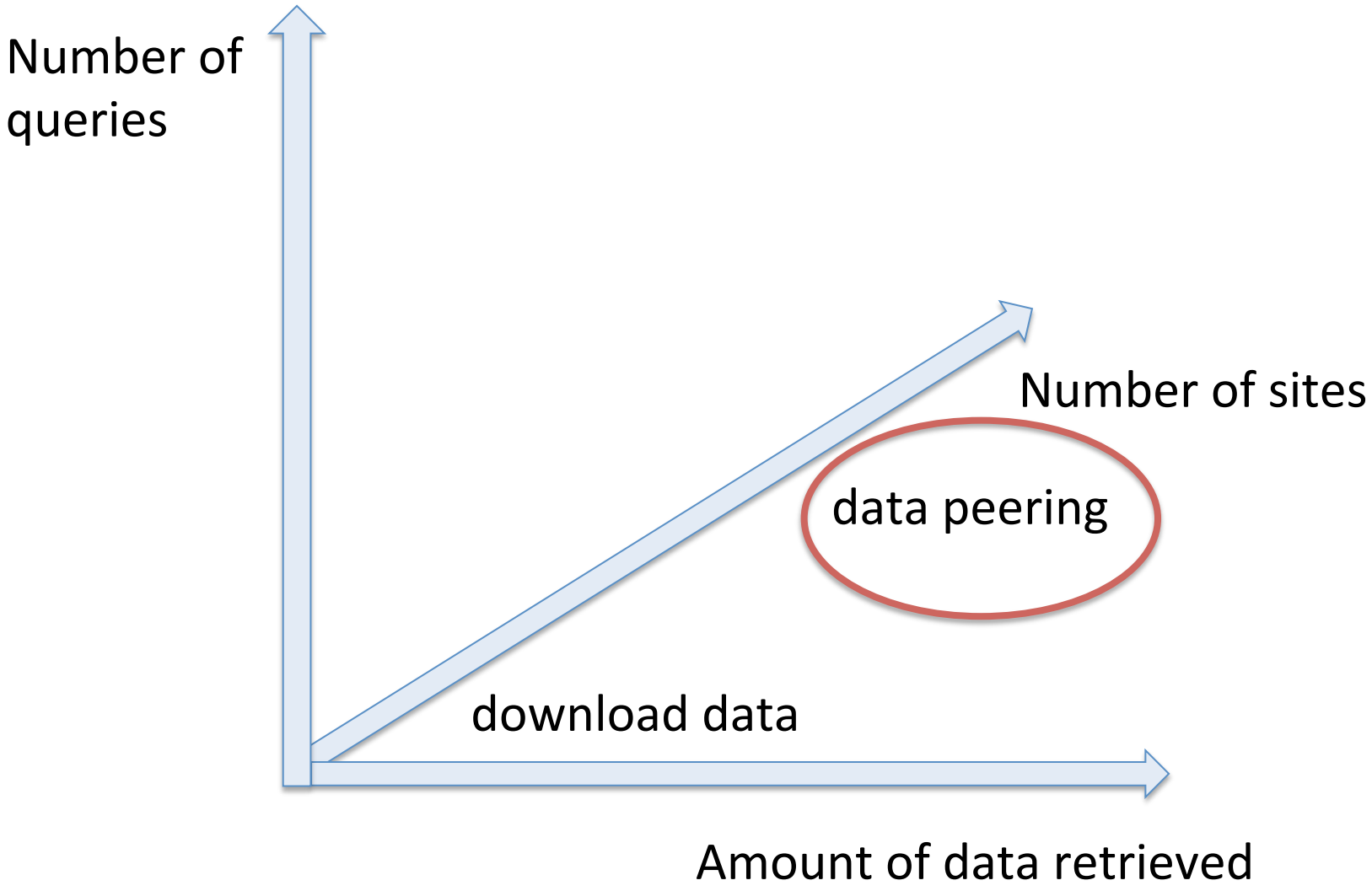
Many ISPs have multihoming where they connect to more than one upstream provider for diversity and reach.

Customers can connect to any ISP.

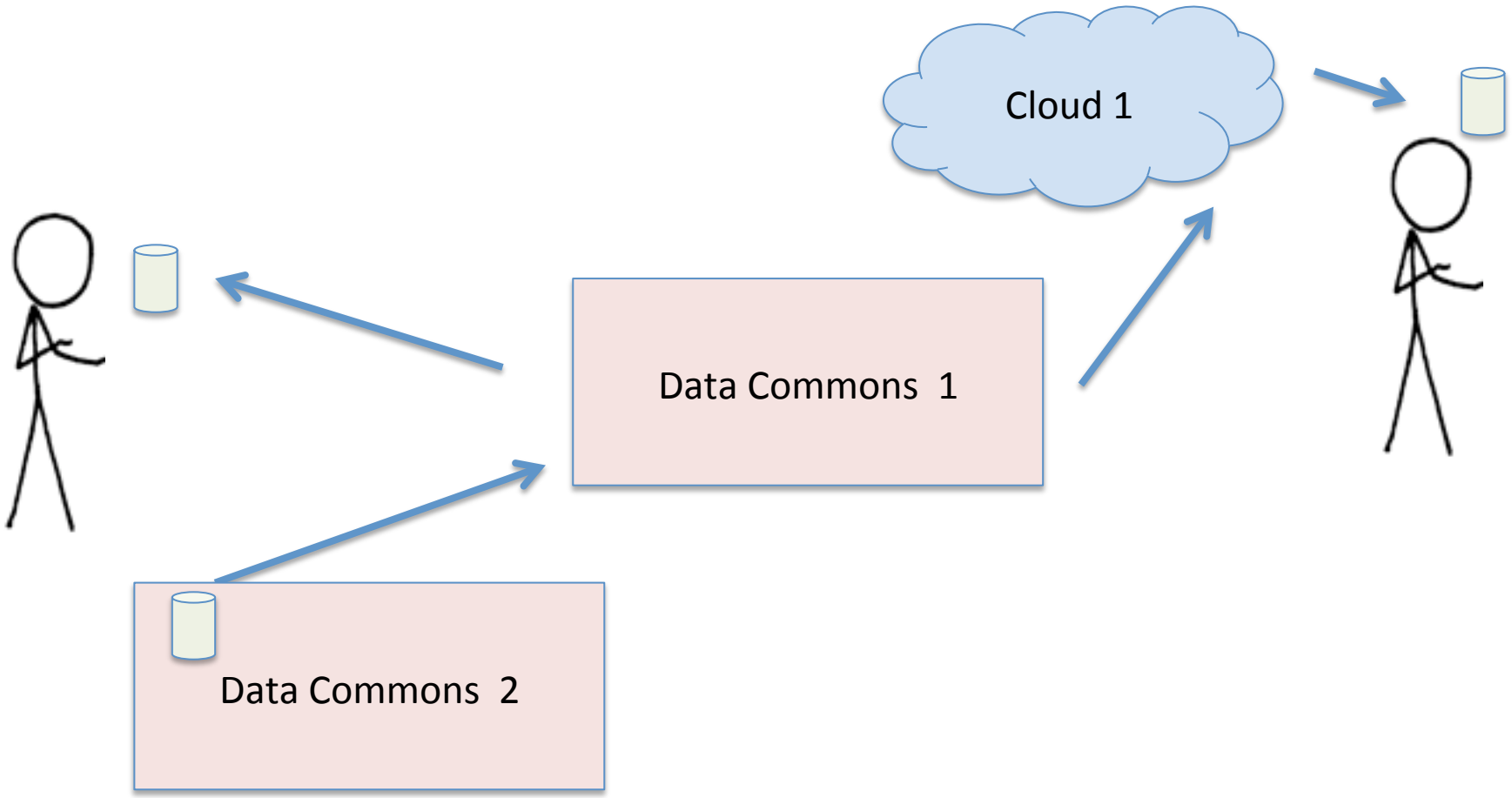


Source: IDC, 2006

Tier 1 ISPs “Created” the Internet


Number of queries
Number of sites
data peering
download data
Amount of data retrieved

Data Peering



- Tier 1 Commons exchange data for the research community at no charge.

Three Requirements

Two Research Data Commons with a Tier 1 data peering relationship agree as follows:

1. To transfer research data between them at no cost.
2. To peer with at least two other Tier 1 Research Data Commons at 10 Gbps or higher.
3. To support Digital IDs (of a form to be determined by mutual agreement) so that a researcher using infrastructure associated with one Tier 1 Research Data Commons can access data transparently from any of the Tier 1 Research Data Commons that hold the desired data.

5. Five Challenges for Data Commons

The 5P Challenges

- **P**ermanent objects with Digital IDs
- Cyber **P**ods with scalable storage and analytics
- Data **P**eering
- **P**ortable data
- Support for **P**ay for compute

Challenge 1: Permanent Secure Objects

- How do I assign Digital IDs and key metadata to "controlled access" data objects and collections of data objects to support distributed computation of large datasets by communities of researchers?
  - Metadata may be both public and controlled access
  - Objects must be secure
- Think of this as a "dns for data."
- The test: One commons serving the earth science community can transfer 1 PB of data files to another commons and no data scientist needs to change their code

Challenge 2: Cyber Pods and Datapods

- How can I add a rack of computing/storage/networking equipment to a *cyber pod* (that has a manifest) so that
  - After attaching to power
  - After attaching to network
  - No other manual configuration is required
  - The data services can make use of the additional infrastructure
  - The compute services can make use of the additional infrastructure
- In other words, we need an open source software stack that scales to cyberpods and data analysis that scales to datapods.

Challenge 3: Data Peering

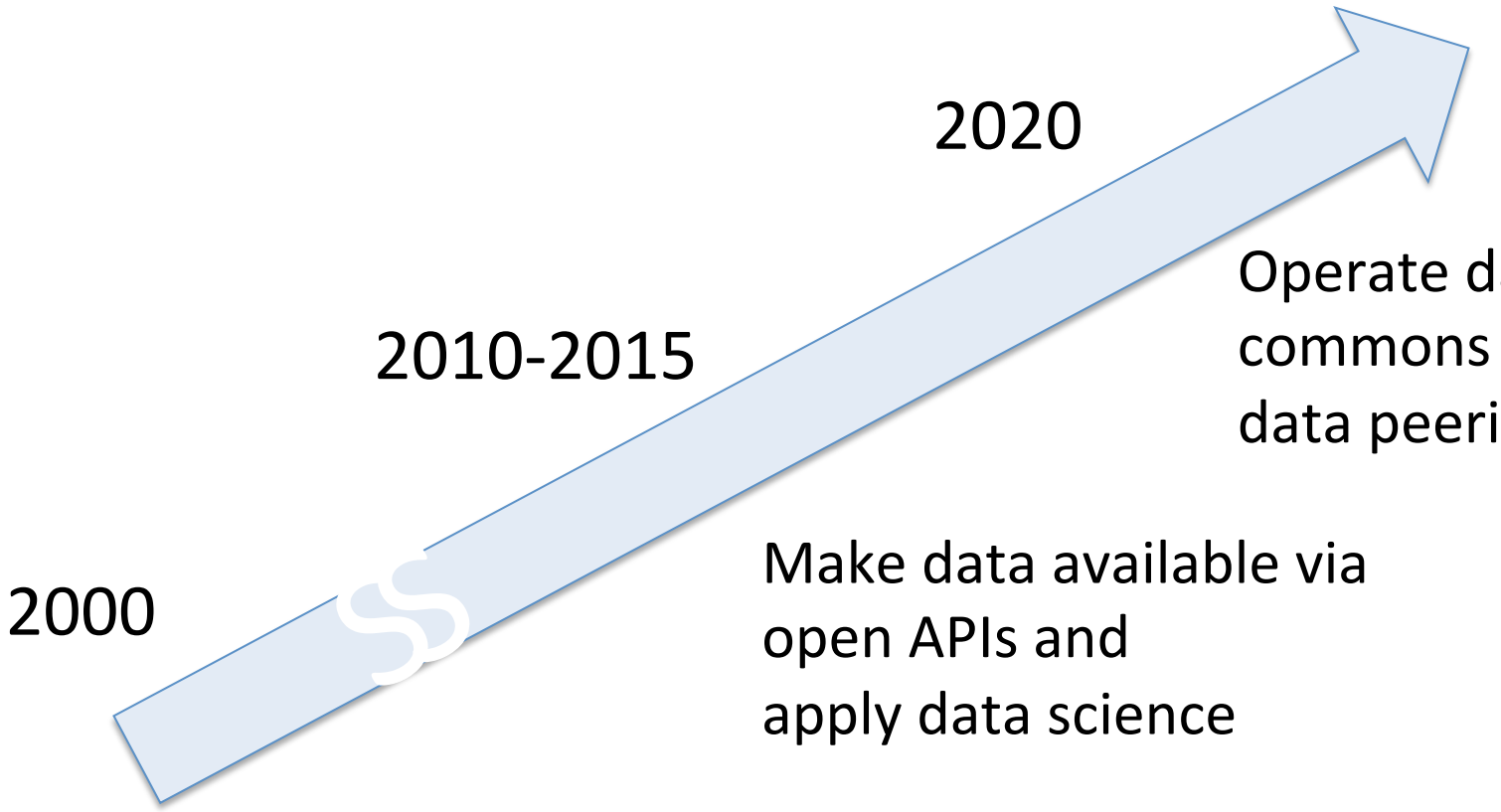
- How can a critical mass of data commons support data peering so that a research at one of the commons can transparently access data managed by one of the other commons
  - We need to access data independent of where it is stored
  - "Tier 1 data commons" need to pass community managed data at no cost
  - We need to be able to transport large data efficiently "end to end" between commons

Challenge 4: Data Portability

- We need an “Indigo Button” to move our data between two commons that peer.

Challenge 5: Pay for Compute Challenges – Low Cost Data Integration

- Commons should support a "free storage for research data, pay for compute," model perhaps with "chits" available to researchers.
- Today, we by and large integrate data with graduate students and technical staff
- How can two datasets from two different commons be "joined" at "low cost"
  - Linked data
  - Controlled Vocabularies
  - Dataspaces
  - Universal Correlation Keys
  - Statistical methods


2020
Operate data commons & support data peering
2010-2015
Make data available via open APIs and apply data science
2000
Collect data and distribute files via DAAC and apply data mining

Questions?

For more information:

rgrossman.com

@bobgrossman